LEVEL

REPLY TO SHAFER: LINDLEY'S PARADOX

D. V. Lindley

AD A100615

Mathematics Research Center

University of Wisconsin—Madison

610 Walnut Street

Madison, Wisconsin 53706

May 1981

(Received April 3, 1981)
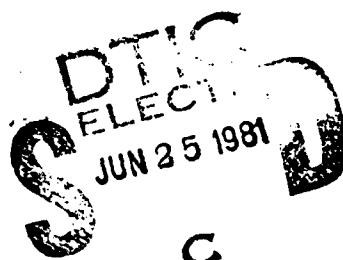
DTIC
ELECT
JUN 2 5 1981
C

DTIC FILE COPY

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

REPLY TO SHAFER: LINDLEY'S PARADOX

D. V. LINDLEY

Technical Summary Report #2212

May 1981

ABSTRACT

This paper is an invited reply to Shafer, G. (1981). Lindley's paradox,

Journal American Statistical Association (to appear).

AMS (MOS) Subject Classifications: 62F03, 62P99

Key Words: Belief functions, scoring rules, Bayesian hypothesis test, glass

fragments, density estimation.

Work Unit Number 4 - Statistics and Probability
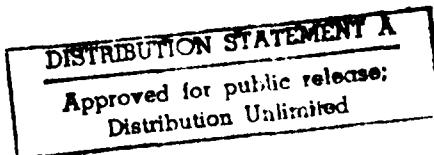
SIGNIFICANCE AND EXPLANATION

Forcible entry has been made to a building, by breaking a window, and a crime committed. A suspect is later found to have fragments of window glass adhering to his clothing. Measurements of the refractive indices of the glass at the scene of the crime and on the suspect's clothing are made and found to be similar. What evidence is there that the clothing glass came from the window?

The topic has been much discussed in the forensic science literature. I (Biometrika, 64, 207-213 (1977)) gave a solution. This has been criticized by Shafer in a paper to appear in the J. Amer. Statist. Assoc. (1981). This report is a reply to Shafer prepared at the request of the editor.

The problem is of general importance in addressing two fundamental issues: how should we measure the strength of an apparent coincidence (between the two types of glass) and no unusual event should be considered without reference to alternatives (what is the usual value for the refractive index of window glass?).

REPLY TO SHAFER: LINDLEY'S PARADOX

D. V. Lindley

The supporter of a theory should welcome good criticism: and I know of no better critic of the Bayesian viewpoint than Shafer. If the theory survives the criticism, then it is enhanced the more the better the critique. In my view, Bayesian ideas come out of Shafer's analysis rather well.

1. <u>Reliability of evidence</u>. It is not always recognized that only the relevant probability matters: whether that probability is based on strong or weak evidence is immaterial. Shafer is wrong when he says "he ought also to weigh the reliability of the evidence". Consider the following example. An urn contains a large number of balls each of which is coloured either red or black, one of them is to be drawn at random and a prize awarded if the ball is red. Contrast two situations. In the first the proportion of red balls is known to be $\frac{1}{2}$. In the second the proportion p is unknown but is described by a probability density $f(p)$ with mean $\frac{1}{2}$. As far as the prize is concerned the relevant probability is that of a red ball being drawn, which is $\frac{1}{2}$ in both situations. The fact that the knowledge of p is less reliable in the second case is irrelevant. Tversky (1974) reports that in a choice between the two situations subjects incoherently prefer the first. Shafer appears to share their view when he discounts the histogram evidence, for only the probability of guilt is relevant.

The reason for the confusion is that the irrelevant aspects can become relevant if the problem is changed and a different probability required. To see this modify the examples to where <u>two</u> balls are to be drawn and the prize

awarded if they are of the same colour. The relevant probability for a given p is $p^2 + (1-p)^2$. This is $\frac{1}{2}$ in the first case but $\int [p^2 + (1-p)^2] f(p) dp$ in the second. This is easily evaluated to give $\frac{1}{2} + 2\sigma^2$ where σ^2 is the variance of p. Now the situations are distinguishable. There are similarly aspects of the histogram evidence that would be relevant for some questions, but for the question of guilt the strength of that evidence does not matter any more than did that about p in the example.

2. <u>Behavioural assessment</u>. In discounting the histogram evidence, Shafer uses a rate α. What does this number mean? He argues that a behavioural interpretation is not necessary but other than by behaviour how can we understand α? Bayesian arguments are firmly based on behaviour. Shafer claims that "Bayesian theory uses canonical examples where the truth is generated according to known chances". This is a possibility, but not the only one. Thus Ramsey's (1931) canonical form is "an ethically neutral proposition of degree of belief $\frac{1}{2}$". An event is ethically neutral if you do not mind whether it is true or false. If has degree of belief $\frac{1}{2}$ if you are indifferent between receiving a prize contingent on the truth, or on the falsehood of the event. No chance element enters here. Or take de Finetti's (1974) scoring rule in which belief a for an event A attracts a penalty score $(A-a)^2$, where A also denotes the indicator function of A. Here no canonical form is used. (Incidently this method is available for any scoring rule and not just the quadratic.) If such a rule is applied to belief functions, in which A and 1 - A may have beliefs that add to less than one, then these beliefs will never attract a smaller score than those based on probabilities: Lindley (1981). Shafer's procedure is inadmissible for any scoring rule.

3. <u>Comparisons of small probabilities</u>. Shafer rightly points out that in the forensic case the Bayesian method compares two probabilities (of the data on the null and alternative hypotheses) both of which are typically small, and he suggests this is unsatisfactory. This is not so: the comparison of small probabilities is the usual situation because most things that happen to us have low probability; we go through life experiencing rare events. You are giving a lecture and collect a list of the students' names. Afterwards you look at the list and see that the probability of those names is very low. If there were only 10 possible names and 8 students, it is 10^{-8} (and this includes the case where all the names are the same). We pass the coincidence by unless we can think of another hypothesis that increases the small probability substantially. It is a basic, important principle of life that we should only judge things in comparison with other things. Neyman and Pearson taught us this in statistics: compare $p(x|H)$ with $p(x|H')$. In the forensic case any measurement on the suspect has low probability - indeed, in the ultimate, perfectly accurate, mathematical fiction, it has probability zero. It is therefore appropriate that two low values should be compared.

4. <u>The soundness of legal arguments</u>. There is a tacit assumption in some philosophical and statistical writing about legal matters that the law is right. Cohen (1977) makes this rather explicit in his book. And Shafer seems to support the view when he argues that defense counsel would attack the Bayesian argument on the grounds discussed in my section 3. To this my answer is that we should not accept legal arguments uncritically but compare them with those suggested by the coherent approach to see the merits of each. When we do this we see that the essentially destructive nature of arguments used by counsel is unsatisfactory because it does not involve consideration of alternatives. Finkelstein (1978) makes the sensible suggestion that a defense

counsel should be required to produce alternative, positive proposals for the prosecution to criticize. Any competent lawyer could destroy any scientific theory.

On another legal matter Shafer suggests that weighing of evidence is not allowed by witnesses. Here the legal and Bayesian arguments do not conflict My suggestion to the forensic scientist is that he should give the probabilities of the data (evidence) both on the supposition of guilt and on that of innocence. The jury can then process these values by taking their ratio and multiplying by the odds without the forensic evidence; thereby performing the weighting. In general, it is the task of the witness to provide all or part of the likelihood. The expert should not do what Shafer suggests and testify that there are "very great odds for the hypothesis" since he has no right to speak to the prior probability. This was the basic mistake made in the Collins case where the likelihood ratio was fairly sound and large, but the final odds were only modest because the prior odds were so small.

5. "Lumpiness". Before tackling the specific issue let me make a general point. Probability is a function of two arguments: the event being assessed, A, and the conditions under which the assessment is being made, H. We write $p(A|H)$. Probability is often taught as if it were part of measure theory. This ignores much of the beauty and importance of the subject, for it is a measure only as a function of A - not as a function of H. The relevance of this general remark here is that my paper had a specific H, namely the histogram of Figure B. As other evidence is accumulated, H changes and so may the probability. Shafer, just as a lawyer would, brings in additional evidence; and it is right that they should do so. For example the

-4-

other histograms tell us more about window glass. But it is unreasonable to criticize p(A|H) because it is not p(A|H'). In science, and in law, we should include all in H that we reasonably and economically can.

Now for the lumpiness. Let me make the assumption that the measurements that lead to the histograms are made with the same precision as those in the trial evidence. Then the quantity required is

$$L_1 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp[-\frac{(y-\theta)^2}{2\sigma^2}] \pi_1(\theta)d\theta$$

where $\pi_1(\theta)$, or more correctly, $\pi_1(\theta|D]$, is the probability of θ given the histogram evidence D. This is equal to $p(y|y_1,y_2,\ldots,y_n)$ where $D = (y_1,y_2,\ldots y_n)$ and all the y's are judged exchangeable. So all we are saying is that y is just like the fire data $\{y_i\}$. The question therefore reduces essentially to evaluating the density function of the y's and the statistical literature is rich in useful methods. (I did a rather "sloppy" job here because my concern in the paper was to emphasize other points.) All these methods use smoothing and the better ones estimate the smoothing hyper-parameter. If there is additional evidence about the smoothing then this could be incorporated into the prior. Actually it is clear that L_1 is not much affected by lumpiness. For example, L_1 will scarcely be altered if 30 values are all at y or 30 values are spread over $y \pm \sigma$, for L_1 is a smoothed version of $\pi_1(\theta)$, smoothed by the error in y.

There is a point where the lumpiness does matter. In my paper the assumption was made that if the glass on the suspect's clothing did truly match θ_0 then he was guilty. But if there are lumps, this may not be so for

-5-

there may be several windows with index θ_0 and all that the evidence could show is that the glass came from one of these, not necessarily from the window at the scene of the crime.

6. **Miscellaneous comments**. Seheult (1978) and Grove (1980) have both commented on my paper and their criticism is worth studying although neither make reference to the fact that their proposals are incoherent.

It was assumed in that paper that the glass was window and not, for example, bottle glass. My understanding was that it was possible to distinguish between the various broad types of glass.

A problem that does need analysis is that suggested by Shafer in his second comment in 5.3 when more than one piece of window glass is found on the suspect. There are several possibilities: none of the glass came from the broken window, only one piece did, two pieces did, and so on. It becomes a little messy to compare all the possibilities.

Is Shafer correct when he refers to the precision of an average? Is he not confusing precision with accuracy? Precision may be measured by the inverse of the variance: accuracy by the inverse of the mean-square error. Because scientific measurements typically contain unknown and undetected biases, precision can increase without limit but not accuracy. Statisticians with their emphasis on standard errors that ignore the bias have confused the issue in some scientific experimentation because the error they quote is substantially less than the true error.

There is one unsatisfactory feature of the Bayesian analysis that Shafer does not mention. It is sensitive to the error distribution. For example, if $(Y-\theta)/\sigma$ has a t-distribution on 5 degrees of freedom, then at $Y-\theta = 2\sigma$ the likelihood is 0.171 times its value at $Y = \theta$, compared with 0.135

for the normal: at 4σ the values are 1.35×10^{-2} and 3.35×10^{-4}, respectively. We need more information about the tails of the error distribution.

There is room for improvement in the details of the Bayesian analysis of forensic data but the basic principles seem untouched by the criticism offered in the paper.

REFERENCES

COHEN, L. JONATHAN (1977). <u>The Probable and the Provable</u>. Oxford: Clarendon Press.

DE FINETTI, B. (1974). <u>Theory of Probability</u>, Vol. 1. New York: Wiley.

FINKELSTEIN, MICHAEL O. (1978). <u>Quantitative Methods in Law</u>. New York: The Free Press.

GROVE, D. M. (1980). The interpretation of forensic evidence using a likelihood ratio. <u>Biometrika</u>, 67, 243-246.

LINDLEY, DENNIS V. (1981). Scoring rules and the inevitability of probability. <u>Operations Research Center</u> 81-1. Berkeley: University of California.

RAMSEY, FRANK PLUMPTON (1931). <u>The Foundations of Mathematics and Other Essays</u>. London: Kegan, Paul, Trench, Trubner.

SEHEULT, ALLAN (1978). On a problem in forensic science. <u>Biometrika</u> 65, 646-648

TVERSKY, AMOS (1974). Assessing uncertainty. <u>J. R. Statist. Soc B.</u> 36, 148-159.

DVL/jvs

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2212	2. GOVT ACCESSION NO. AD-A100 615	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) REPLY TO SHAFER: LINDLEY'S PARADOX		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) D. V. Lindley		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of 610 Walnut Street Wisconsin Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE May 1981
		13. NUMBER OF PAGES 8
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Belief functions, scoring rules, Bayesian hypothesis test, glass fragments, density estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper is an invited reply to Shafer, G. (1981). Lindley's paradox, Journal American Statistical Association (to appear).

DD FORM 1473   EDITION OF 1 NOV 65 IS OBSOLETE
1 JAN 73

DATE
ILMED

−8